XSearch: A System for Searching and Interrelating NASA Mission Operations Data

Richard M. Keller[1], Shawn R. Wolfe[2], and May N. Windrem[3]
NASA Ames Research Center, Moffett Field, CA, 94035-1000

Daniel C. Berrios[4]
University of California Santa Cruz, Moffett Field, CA, 94035-1000

To perform their duties, flight controllers for the Space Shuttle and International Space Station routinely need to locate specific operations documents and records from among tens of thousands available. Relevant information, in the form of team notes, console logs, action requests, anomaly reports, and flight procedures, is stored in heterogeneous databases and accessed using varying tools with differing interfaces. Users must consult these different systems and manually integrate results together to get a comprehensive view of relevant information. To improve access to flight control information, we have designed and built an application, XSearch, that integrates data from disparate flight operations databases. Through a common search interface, Mission Control personnel using XSearch can issue a single search query and simultaneously interrogate multiple mission operations data sources. The initial version of XSearch is planned for deployment in NASA's Mission Control Center in mid-2008, and will allow Shuttle and Station flight controllers to search simultaneously across three key mission operations data sources: the ChitS system (used to store mission action requests), the Flight Notes system (used to store internal flight control team communications), and the Anomaly Reporting System. These systems store historical data back to 2002, and contain over 100,000 records in total. XSearch users can perform full-text searches on key text fields (e.g. title, problem description, action request, etc.) and view integrated results across these data sources. In addition to conducting search, the system provides two other important capabilities that are intended to contextualize search results: detection of cross-references and detection of textually similar records. Identifying records that are either cross-referenced by, or similar to, a given search result enables flight controllers to recover key information about the operational context associated with that result. The goal of contextualization is to facilitate safer and more effective mission operations decision-making through enhanced situation awareness. To detect embedded cross-references within results, XSearch parses text fields found in the results using a set of syntactic patterns that identify citations (e.g., patterns that detect controlled document or record identifiers routinely used by authors). Using this technique, XSearch can identify both "outbound" and "inbound" references. Outbound references point "out" from a specific chit, flight note, or anomaly to other records; "inbound" references point "in" to the specific item from other records. To detect records that appear similar to a given search result, XSearch calculates and ranks the textual similarity between the result and all other records in the corpus; those ranked highest are displayed to the flight controller. Similarity detection is computed using a standard cosine-based vector space information retrieval method weighted by term

[1] Computer Scientist, Intelligent Systems Division, Mail Stop 269-2, non-member.
[2] Computer Scientist, Intelligent Systems Division, Mail Stop 269-2, AIAA member.
[3] Computer Scientist, Intelligent Systems Division, Mail Stop 269-2, non-member.
[4] Computer Scientist, University Affiliated Research Center, Mail Stop 269-2, non-member.

frequency (TF) multiplied by inverse document frequency (IDF). This technique identifies items that may be relevant to a given record based on similarity in terms, even though the items are not explicitly cited. The initial version of XSearch provides a framework for the development of more sophisticated and powerful information correlation features. Our long-term strategy is to expand the information extraction methods we employ to recover many other types of cross-record linkages.

I. Introduction

FLIGHT controllers for NASA's International Space Station (ISS) and Space Shuttle have demanding jobs that require continual study of a large quantity of technical information about the systems and processes they monitor and control. As there is far too much information to commit to memory, flight controllers typically access a variety of computer systems to locate information relevant to their tasks, including operations manuals, flight procedures, flight rules, action requests, anomaly reports, and console logs. They must be adept at locating necessary information in a timely fashion in order to respond to contingencies that may arise rapidly during a mission. Before acting on the information, they must also understand the operational context at the time the information was generated. Without adequate situational awareness, the flight controller might misinterpret information, draw incorrect inferences, and make poor decisions.

Although the importance of maintaining situational awareness is widely recognized, both locating information and interpreting that information in the proper operational context can be challenging in the current Mission Control Center (MCC) environment at NASA Johnson Space Center (JSC). The environment hosts a patchwork of separate information management systems, each with separate and distinct search and access capabilities. This is not surprising considering the 20-year time span of the Shuttle and ISS programs and the costs involved in retrofitting old systems to interoperate with newer ones. To consolidate access to the myriad of information resources required for flight control, JSC engineers have built a unifying web portal. However, the portal is only a thin connecting interface layer that sits atop a set of disconnected systems. Flight controllers must still search for information across multiple resources by querying each system separately. Then they must manually collate the results across multiple displays – a time-consuming and error-prone process. Differences in search functionality across the various systems can frustrate flight controllers further.

Interpreting information retrieved via search is also a challenge for flight controllers – particularly with respect to understanding the relevant operational context. During operations, flight controllers and mission support personnel create numerous documents and electronic records as routine by-products of their activities. This set of information forms the basis for understanding the operational context at a given point in time. Unfortunately, the relationships among these by-products are not systematically captured in a way that enables rapid retrieval of the operational context at a subsequent point. The lack of a 'breadcrumb trail' makes support for situational awareness quite difficult. For example, suppose a flight controller has retrieved an anomaly report in response to a search for a specific type of sensor failure. The flight controller should also have access to related information that provides a context for understanding the anomaly, including the conditions under which it occurred and the actions taken to remedy the problem. If a software workaround was developed in response to the anomaly, the flight controller should be able to access that information easily. However, the anomaly record may not actually contain a pointer to the software workaround procedure, which is stored in a different database. The connection between these two records may only be captured as part of an unstructured text description entered by a flight controller in a log entry. Without an explicit structured representation at the data storage level, these types of connections cannot be displayed automatically to users. Instead, flight controllers must rely on their ability to scan text and detect references to key documents whose content may impact their interpretation and decisions. This can be problematic when the text description is lengthy and the flight controller is under pressure.

This paper describes a system, XSearch, which is being deployed within the MCC environment to address the key information access and retrieval challenges described above. XSearch enhances flight controllers' ability to search across distributed mission operations information and also reconstructs aspects of the operational context necessary for them to interpret and apply that information correctly. The paper is organizes as follows: Section II elaborates on the notion of information context and describes our approach to automated context recovery; Section III presents the XSearch system and its methods; Section IV describes related work; Section V discusses implementation status, future directions, implementation challenges, and observations; and Section VI concludes.

II. Information Context and its Recovery

For the purposes of this paper, we define *information context* in terms of a corpus of information that may include operational documents, database records, images, spreadsheets, web content, etc.:

Definition 1: Given a corpus of information, the *information context* associated with a given operational event or issue is the subset of the corpus that is *operationally relevant* to understanding that event or issue.

This notion of information context is not precise because the assessment of operational relevance is subjective, and varies based on the evaluator's knowledge, experience, and judgment. Nevertheless, the notion of information context provides a starting point for conceptualizing the set of documents that are necessary to support flight controller situational awareness. Because the information context is not captured explicitly in the MCC environment, yet is important for situational awareness, we implemented techniques in XSearch to recover the information context and display it to the user. The recovery process cannot be precise or complete due to the lack of adequate information capture at the time the operational events or issues occurred. Nonetheless, we believe that partial (and even potentially flawed) reconstruction will be superior to no information context recovery at all.

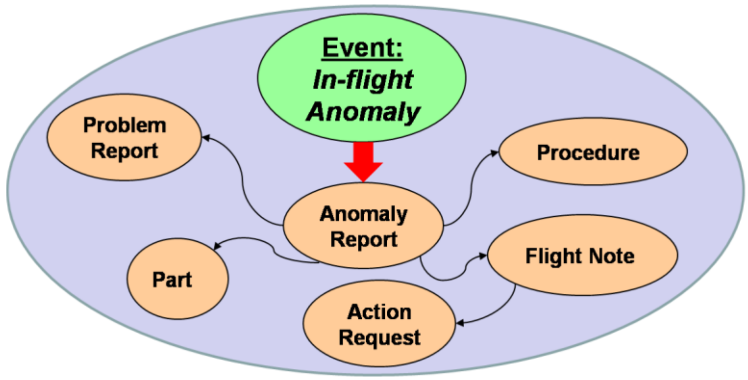


Figure 1. Information context associated with an in-flight anomaly. *Documents, notes, and records pertaining to the detection, work-around, and permanent resolution of the in-flight anomaly are part of its information context.*

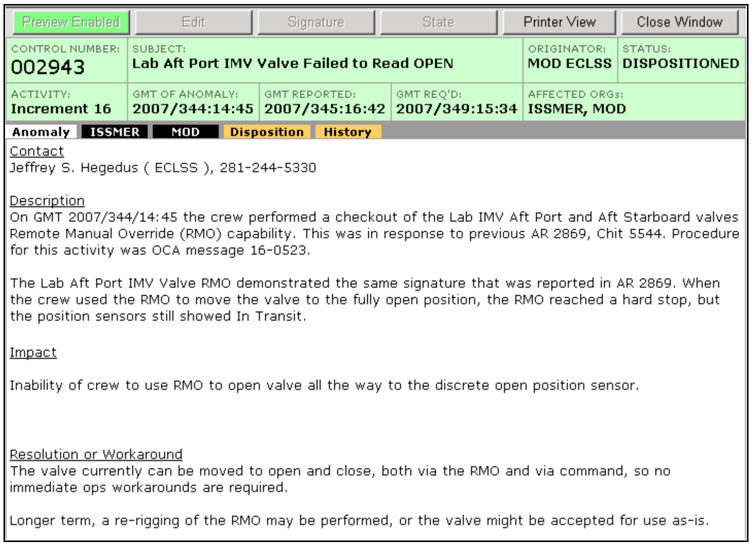
To illustrate the notion of information context, Figure 1 depicts the set of information associated with an in-flight anomaly event. One clearly relevant piece of information is an *anomaly report* summarizing the anomalous operational symptoms. In the MCC, each in-flight anomaly is documented by a single anomaly report, which can be considered as a proxy for the actual event. However, the anomaly report is only the tip of the 'information iceberg' relating to the event. A number of other documents also contain information necessary to form a more complete picture of the anomaly event, for example: a *problem report*, detailing the engineering analysis of factors that caused the anomaly; a *part record*, providing a description and schematics for a failed part involved in the anomaly; a *flight procedure* that was incorrectly executed leading up to the part failure and ensuing anomaly; a *flight note* describing actions taken by the flight control team to mitigate impacts and deal with the anomaly; and an *action request* requesting modification of the flight procedure to prevent future problems.

Determining the information context using Definition 1 would be difficult to accomplish in the MCC environment. First, no master catalog of operational issues or events is maintained by the MCC. The occurrence of events or issues can be inferred only through indirect evidence. For example, an event or issue might be detected by the presence of text references in documents or database fields. The second difficulty with Definition 1 lies in determining a suitable operational relevance criterion. Without a representation of operational issues or events, it is difficult to develop a computational method to assess operational relevance.

We circumvent these problems with the formulation of Definition 1′, which approximates Definition 1 by defining information context in terms of proxy information objects rather than operational events or issues, themselves. As discussed above and illustrated in Figure 1, in some cases an information object can serve as a proxy for an operational event or issue.

Definition 1′: Given a corpus of information, the *information context* associated with a given information object (e.g., a document, database record, image, spreadsheet, web page, etc.) is the subset of information in the corpus that *references* or *is referenced by* the object.

The intuition behind this revised definition is that information referenced by a proxy object might be operationally relevant to the proxied event or issue; a similar argument can be made for information that references a proxy object. For example, consider Anomaly Report 002943 in Figure 2. The text of this report cites another anomaly report (AR 2869), an action request (Chit 5544), and an OCA message (16-0523) by making direct references to these records. Using Definition 1′, these three information objects would be considered part of the information context associated with Anomaly Report 002943. If other chits, flight notes, or anomaly reports reference Anomaly Report 002943, we would include those in its information context, as well. Note that the modified definition is heuristic – only a subset of the objects included might be operationally relevant information. For example, the anomaly report in Figure 2 might have contained a URL link to the anomaly reporting system home page. Although explicit, this web page reference would not have been operationally relevant to the specific in-flight anomaly at issue.



Preview Enabled | Edit | Signature | State | Printer View | Close Window

CONTROL NUMBER:	SUBJECT:			ORIGINATOR:	STATUS:
002943	Lab Aft Port IMV Valve Failed to Read OPEN			MOD ECLSS	DISPOSITIONED
ACTIVITY:	GMT OF ANOMALY:	GMT REPORTED:	GMT REQ'D:	AFFECTED ORGs:	
Increment 16	2007/344:14:45	2007/345:16:42	2007/349:15:34	ISSMER, MOD	

Anomaly | ISSMER | MOD | Disposition | History

Contact
Jeffrey S. Hegedus (ECLSS), 281-244-5330

Description
On GMT 2007/344/14:45 the crew performed a checkout of the Lab IMV Aft Port and Aft Starboard valves Remote Manual Override (RMO) capability. This was in response to previous AR 2869, Chit 5544. Procedure for this activity was OCA message 16-0523.

The Lab Aft Port IMV Valve RMO demonstrated the same signature that was reported in AR 2869. When the crew used the RMO to move the valve to the fully open position, the RMO reached a hard stop, but the position sensors still showed In Transit.

Impact

Inability of crew to use RMO to open valve all the way to the discrete open position sensor.

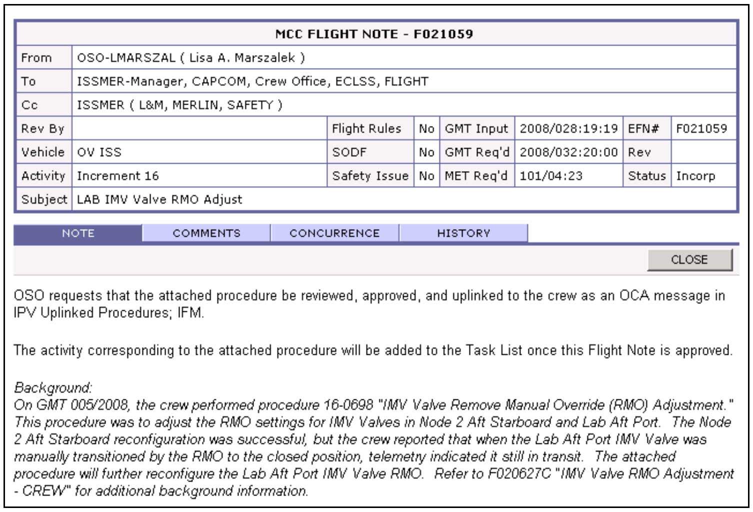
Resolution or Workaround
The valve currently can be moved to open and close, both via the RMO and via command, so no immediate ops workarounds are required.

Longer term, a re-rigging of the RMO may be performed, or the valve might be accepted for use as-is.

Figure 2. Anomaly report exhibiting cross-references. *The text of this anomaly report mentions other related documents that provide context for understanding the in-flight anomaly.*

Definition 1′ was formulated based on the assumption that identifying references is the key to recovering the information context. We define three types of references that may be present in the text of an information object: explicit references, inferred references, and implicit references. *Explicit references* include links from the proxy information object to other objects based on the inclusion of unique identifiers. In Figure 2, the references to the anomaly report, chit, and OCA message all make use of unique record identifiers to identify the information object being cited. But such references are only one kind of explicit reference. Other types of explicit references include those generated automatically by an underlying software application. For example, if a proxy information object is under version control, the underlying version control system automatically might create an explicit link to previous versions of the object. Another type of explicit references is the type made when a user fills in a link entry field in a form. For example, a user interface might allow users to link a database entry to one or more related records by selecting from a pre-populated list.

Inferred references are indirect references that require background knowledge and deductive capabilities to detect. An example of a textual reference that must be inferred is 'the crew assignment procedure developed by George Simpson during Shuttle Increment 06.' The procedure record is not referenced using an explicit identifier, and resolving the reference requires application of knowledge (e.g., that the author of a procedure and the date of its establishment are recorded in the procedures database) and inference (e.g., methods to determine which procedures



MCC FLIGHT NOTE - F021059

From	OSO-LMARSZAL (Lisa A. Marszalek)						
To	ISSMER-Manager, CAPCOM, Crew Office, ECLSS, FLIGHT						
Cc	ISSMER (L&M, MERLIN, SAFETY)						
Rev By		Flight Rules	No	GMT Input	2008/028:19:19	EFN#	F021059
Vehicle	OV ISS	SODF	No	GMT Req'd	2008/032:20:00	Rev	
Activity	Increment 16	Safety Issue	No	MET Req'd	101/04:23	Status	Incorp
Subject	LAB IMV Valve RMO Adjust						

NOTE | COMMENTS | CONCURRENCE | HISTORY

CLOSE

OSO requests that the attached procedure be reviewed, approved, and uplinked to the crew as an OCA message in IPV Uplinked Procedures; IFM.

The activity corresponding to the attached procedure will be added to the Task List once this Flight Note is approved.

Background:
On GMT 005/2008, the crew performed procedure 16-0698 "IMV Valve Remove Manual Override (RMO) Adjustment." This procedure was to adjust the RMO settings for IMV Valves in Node 2 Aft Starboard and Lab Aft Port. The Node 2 Aft Starboard reconfiguration was successful, but the crew reported that when the Lab Aft Port IMV Valve was manually transitioned by the RMO to the closed position, telemetry indicated it still in transit. The attached procedure will further reconfigure the Lab Aft Port IMV Valve RMO. Refer to F020627C "IMV Valve RMO Adjustment - CREW" for additional background information.

Figure 3. Similar flight note. *This note discusses a problem that appears related to the anomaly reported in Figure 2, even though the note is not explicitly referenced in that anomaly.*

pertain principally to crew assignment). Named entity recognition (NER) from the field of information extraction[1] is relevant to resolving inferred references. NER consists of a set of information retrieval methods based upon natural language processing and designed to extract inferred references in textual passages.

Implicit references, in contrast with explicit and inferred references, are not based upon identifying a direct or indirect textual citation within an information object. Instead, the relationship between the information object and the 'referenced' object is established by some other means. For example, an implicit reference between two objects may be inferred based on their *similarity*. Object *A* can be considered to reference object *B* implicitly if *A* is very similar to *B* – according to some definition of similarity. The intuition behind this heuristic is that if *A* and *B* are similar, then *B* might be operationally relevant to *A*. For example, the flight note shown in Figure 3 might be considered similar to the anomaly report in Figure 2 because they share similar terms. Both discuss 'Lab IMV valves' and a faulty 'in-transit' sensor reading. These anomalies are not only textually similar, but also operationally relevant to one another and involve a related problem. Similarity between two textual passages can be determined using a variety of information retrieval techniques[2]. Other notions of similarity may be based on factors other than textual content. For example, the set of information objects generated by a specific set of people during a specific timeframe (e.g., the people staffing flight control positions during a specific operations shift) might be considered similar; or the set of objects with identical information contexts might be considered similar because they all reference the same set of information.

The XSearch system analyzes text to discover both explicit and implicit references, thereby reconstructing the information context. (Note that handling inferred references is outside the scope of the current system.) The next section describes XSearch and the specific methods that were used to accomplish contextualization.

III. XSearch System

XSearch is a Web-based search application that was designed to accomplish two main functions: search integration and search result contextualization. The system permits users to search across multiple MCC databases and display integrated results in a single consolidated listing. Moreover, the results are annotated to indicate whether there are associated contextual cross-references. We discuss search integration and result contextualization in the following subsections.

A. Search Integration

XSearch performs integrated search across three MCC databases:

1. **Chits**: The Chits[*] database contains mission action requests representing formal coordination agreements among various organizational entities charged with Space Station and Shuttle operations (e.g., mission operations, engineering, payload operations, program management, international partners, etc.). A chit generally corresponds to a request made by one organization to one or more others, and documents their responses and the overall agreement reached. A chit can be viewed as a contract that describes how to proceed on an issue that impacts multiple organizations.
2. **Flight Notes**: The Flight Notes database consists of a set of records that document operational status, operations policy, and actions planned by the flight control team. Flight notes support the operations team workflow, allowing the team to communicate and coordinate their activities. For example, one or more specific flight notes may be generated to implement a general agreement made in a chit.
3. **Anomaly Reports**: The Anomaly Report database documents all Space Station or Shuttle anomalies detected by the flight control staff. These anomalies are generally reviewed by the engineering staff, which generates corrective action recommendations for the operations staff.

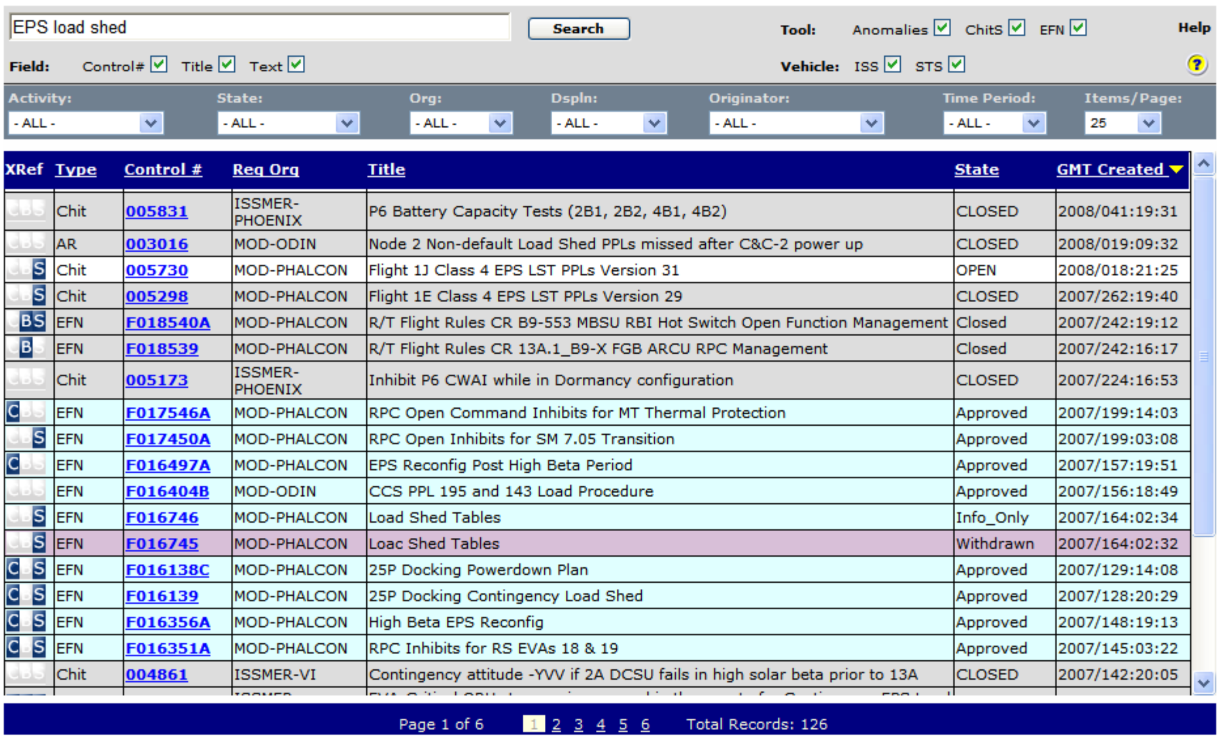
Flight controllers currently access each of these databases using a separate search interface. XSearch executes a cross-database search by querying each of the databases and then presents consolidated results in a single listing (Figure 4). Users enter search queries in the search box at the top of the page and then select basic options for the search. They can select the 'Tools' (i.e., the data sources) to be searched, the 'Fields' within those databases to search, and the 'Vehicles' (i.e. whether to search Space Station records, Shuttle records, or both). Although the databases contain many different fields, we have abstracted away this complexity and provide users with only three choices of fields to search: 'control #', 'title' or 'text'. The control number is the designated record identifier assigned to the chit, flight note, or anomaly report upon initial entry. Flight controllers use these identifying numbers when communicating with each other during the normal course of conducting operations and when authoring written

[*]The term "chit" has its origin in the US Naval Service and was a piece of paper documenting a question and answer between various functions aboard a naval vessel.

records and documentation. Chits, flight notes, and anomaly reports each contains a short informative title field that can be searched. All of the other text entry fields in these records are searched together via the ‘text’ field option. When the user chooses to search the ‘text’ field, XSearch searches all of the primary text fields in the selected databases (except the title, which must be separately chosen).

Beneath the search box in Figure 4 is a set of pull-down lists that can be used to filter the search results based on the *activity* (the mission or mission phase), *state* (the workflow status of the record – whether initiated, in-work, completed, withdrawn, etc.), *org/dscpln* (the organization and flight control discipline that created the record), *originator* (the person originating the record), and *time period* (how many days or months to search back through historical records).

Search query processing differs depending on whether a user quotes the terms they enter into the search box. We modeled the behavior of search on the familiar Google search interface. In particular, if users enter multiple search terms, XSearch retrieves records that each contain all of the terms somewhere within the designated search fields; the terms need not be contiguous or in the same order as entered. If the user wishes to search for an exact phrase, he or she must enter the terms within double quotes, e.g. “EPS load shed”.



XRef	Type	Control #	Req Org	Title	State	GMT Created
	Chit	005831	ISSMER-PHOENIX	P6 Battery Capacity Tests (2B1, 2B2, 4B1, 4B2)	CLOSED	2008/041:19:31
	AR	003016	MOD-ODIN	Node 2 Non-default Load Shed PPLs missed after C&C-2 power up	CLOSED	2008/019:09:32
S	Chit	005730	MOD-PHALCON	Flight 1J Class 4 EPS LST PPLs Version 31	OPEN	2008/018:21:25
S	Chit	005298	MOD-PHALCON	Flight 1E Class 4 EPS LST PPLs Version 29	CLOSED	2007/262:19:40
BS	EFN	F018540A	MOD-PHALCON	R/T Flight Rules CR B9-553 MBSU RBI Hot Switch Open Function Management	Closed	2007/242:19:12
B	EFN	F018539	MOD-PHALCON	R/T Flight Rules CR 13A.1_B9-X FGB ARCU RPC Management	Closed	2007/242:16:17
	Chit	005173	ISSMER-PHOENIX	Inhibit P6 CWAI while in Dormancy configuration	CLOSED	2007/224:16:53
C	EFN	F017546A	MOD-PHALCON	RPC Open Command Inhibits for MT Thermal Protection	Approved	2007/199:14:03
S	EFN	F017450A	MOD-PHALCON	RPC Open Inhibits for SM 7.05 Transition	Approved	2007/199:03:08
C	EFN	F016497A	MOD-PHALCON	EPS Reconfig Post High Beta Period	Approved	2007/157:19:51
	EFN	F016404B	MOD-ODIN	CCS PPL 195 and 143 Load Procedure	Approved	2007/156:18:49
S	EFN	F016746	MOD-PHALCON	Load Shed Tables	Info_Only	2007/164:02:34
S	EFN	F016745	MOD-PHALCON	Loac Shed Tables	Withdrawn	2007/164:02:32
CS	EFN	F016138C	MOD-PHALCON	25P Docking Powerdown Plan	Approved	2007/129:14:08
CS	EFN	F016139	MOD-PHALCON	25P Docking Contingency Load Shed	Approved	2007/128:20:29
CS	EFN	F016356A	MOD-PHALCON	High Beta EPS Reconfig	Approved	2007/148:19:13
CS	EFN	F016351A	MOD-PHALCON	RPC Inhibits for RS EVAs 18 & 19	Approved	2007/145:03:22
	Chit	004861	ISSMER-VI	Contingency attitude -YVV if 2A DCSU fails in high solar beta prior to 13A	CLOSED	2007/142:20:05

Figure 4. XSearch Results Display. *This consolidated search results display incorporates information from multiple operations databases. At the top is a search box for query entry and a set of associated checkboxes to choose search options. Beneath this area is a row of filters that can be selected to narrow the set of search results displayed. Under the filters is a table that lists search results. The results can be sorted by clicking on the column headings. The source of each result is indicated in the ‘Type’ column: Chit indicates a result from the Chits database; AR indicates an Anomaly Report; EFN indicates an Electronic Flight Note. All results contain the search terms somewhere within their titles (under the ‘Title’ column) or text (available by clicking on the hyperlinked identifier in the ‘Control #’ column – see the sample display of anomaly text in* Figure 2 *and flight note text in* Figure 3*).*

B. Search Result Contextualization

As discussed in Section II, the information context consists of the set of information that is operationally relevant to interpreting a given document or record. Contextual information is essential to providing flight controllers with adequate situational awareness knowledge of activities, issues, or events that occurred in the past. The XSearch interface can display three types of contextual cross-references for each search result. ‘*Cites’* cross-references appear

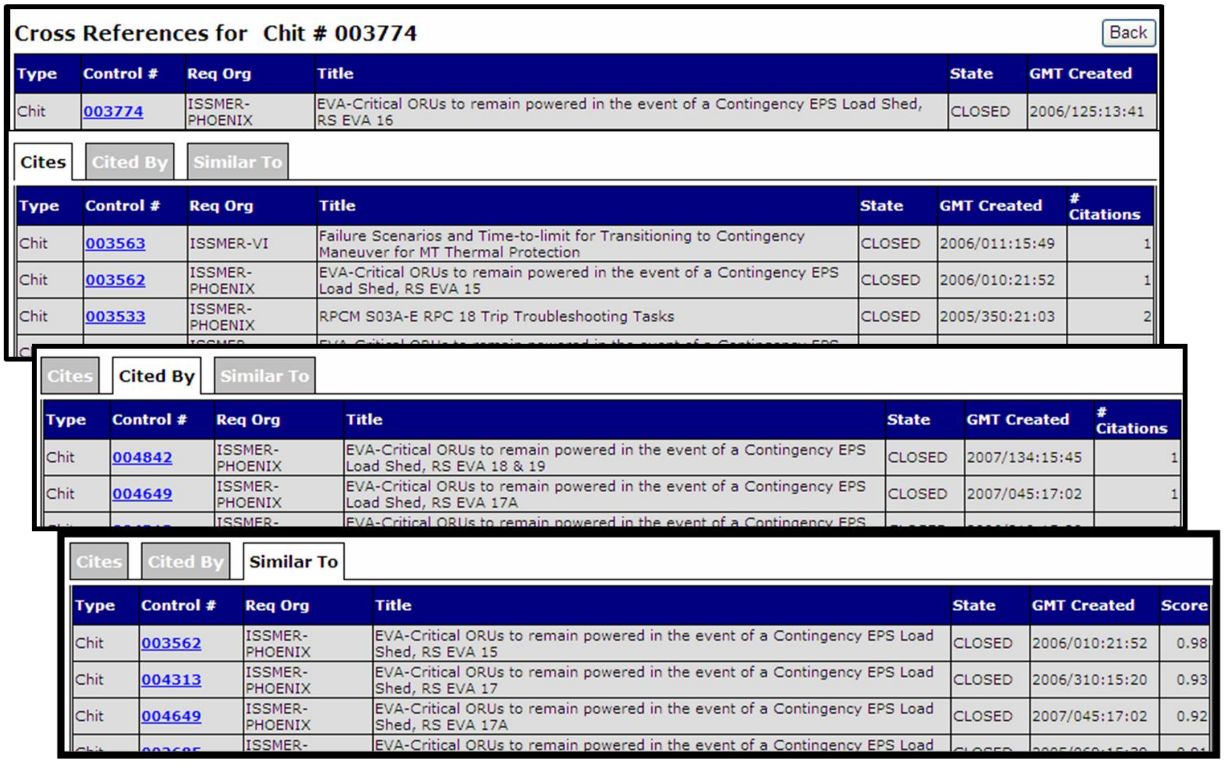


Figure 5. XSearch Cross-Reference Display. *This display presents the three types of cross-references available for the search result listed at top, Chit #003774. In the 'Cites' tab, the first row indicates that Chit #003774 cites Chit #003563 within its text; on the 'Cited By' tab, the first row indicates that Chit #003774 is cited by the text of Chit #004842; on the 'Similar To' tab, the listed chits share many key words in common with Chit #003774. Clicking on the hyperlinked Control # provides access to the cross-reference.*

when a search result explicitly cites another record or document by using its unique identifier (as illustrated by the text in Figure 2); conversely, '*Cited By*' cross-references appear when the search result's identifier is cited by the text of another record or document. '*Similar To*' cross-references appear when the search result contains a significant number of word choices in common with another record or document.

XSearch provides an indication of whether a specific search result has associated cross-references in the 'XRef' column in Figure 4. The presence of a letter icon ('C' for cites; 'B' for cited-by; 'S' for similar-to) means that the system has detected cross-references of this type. Clicking on one of these icons brings up the cross-reference display in Figure 5. Each of the tabs in the display lists a different type of cross-reference. The same sort of information displayed for search results is displayed for each cross-reference. In addition, the 'cites' and 'cited by' tabs contain a column that indicates how many times the cross-reference was cited in the search result. The 'similar-to' tab is sorted according to the 'score' column, which indicates how similar the cross-reference was to the search result on a scale from 0-1.

To produce the display in Figure 5, XSearch constructs and maintains a cross-reference table and a similarity table. The cross-reference table contains information about citations found in operations database records. If record $r1$ contains n citations to record $r2$, then the triple $<r1, r2, n>$ appears as a row in the table. The *cites* and *cited-by* cross-references displayed for a given record can be computed directly by querying this table. The similarity table contains information about the similarity score between each pair of records in the overall corpus. It contains triples of the form $<r1, r2, s>$, where s is the similarity score comparing the textual content of records r1 and r2. (Similarity scores are symmetric, so s also represents the similarity score between r2 and r1.) The *similar-to* cross-references displayed for a given record are obtained by querying the table for matching records above a predetermined similarity threshold (currently set to 0.7) and sorting the results in order of descending similarity score. (Records with scores of 1.0 are most likely exact duplicates of the search result, and are not displayed.)

The cross-reference and similarity tables are created initially by batch processing all records in the chit, flight note, and anomaly report databases. This process takes several hours to complete. The tables are kept up-to-date using a background process that frequently polls the three databases to determine whether any records were added, deleted, or modified by flight controllers during the time elapsed since the previous inquiry. (Flight controllers add or modify records using separate applications that access and update the databases.) If changes are detected, the cross-reference and similarity tables are updated accordingly. The following subsections describe the processing steps required to update the tables when a new record is detected.

1. Detecting Citations

The text of each new record is parsed to detect citation patterns that signify cross-references to other database records. These citation patterns generally take the form of a database identifier followed by a record identifier. For example, the following citations all reference the same record: Flight Note F016356A, EFN F016356A, FN 016356A, flight note 16356A. The parser detects references to multiple records, as well as single citations. For example, three chit citations and one anomaly report citation are parsed from the following sentence: "As noted in Chits 3562, 4649, and 003562, AR 1756 has not yet been resolved." The set of patterns is updated whenever missed citations are discovered by users and reported to the development team. To unify across all the different syntactic forms that flight controllers may use to reference a single record, each citation detected is mapped into a canonical representation and the number of citations is tallied for display in the cross-references table (Figure 5).

2. Computing Similarity

Similarity scores between two records are computed using a cosine measure applied in conjunction with a vector space document representation[3]. In the vector space representation, the textual content of each record is represented as a "bag of words", which tallies the frequency of terms used in a record but does not account for word order, sentence structure, or semantic features of the content. Each record can be represented as an *n*-dimensional vector, where n is the total number of distinct terms occurring in the record corpus. The value of the n^{th} dimension of the record's vector can be assigned a weight, for instance by using the TF-IDF formula[3] (see Eq. (1)), which employs term frequency (TF) and inverse document frequency (IDF). The TF portion increases the weighting by the number of recurrences of a specific term within a document, while the IDF portion reduces the weight if the word is very common – and therefore less informative – within the overall corpus.

Given a corpus of m records containing n distinct terms, the TF-IDF formula for the i^{th} term of the j^{th} record is

$$tfidf(i,j) = w_{ij} * \log \tfrac{m}{\#docs_i} \tag{1}$$

where w_{ij} is the number of occurrences of the i^{th} term in the j^{th} record, and $\#docs_i$ is the number of records that contain the i^{th} term. When the j^{th} record does not contain the i^{th} term, the value of w_{ij} (as well as the *tfidf* function) will be zero.

Application of the TF-IDF formula yields a point in an *n*-dimensional space for each record; we can define the similarity between two records as the cosine of the angle between their corresponding vectors. The more similar the records are to each other, the smaller the angle in between them and the larger the cosine value. The mathematical definition of cosine yields the similarity score formula:

$$\begin{matrix} Similarity \\ Score \end{matrix} = sim(d_j, d_k) = \frac{\sum_{i=1}^{n} \left(tfidf(i,j) \times tfidf(i,k) \right)}{\sqrt{\sum_{i=1}^{n} tfidf(i,j)^2} \sqrt{\sum_{i=1}^{n} tfidf(i,k)^2}} \tag{2}$$

The vector space document model, the TF-IDF weighting formula, and the cosine similarity measure, though somewhat ad-hoc, are surprisingly effective. Variants of the formulae exist, but the overall scheme has withstood the test of time, and is a reasonable starting point for determining record similarity.

IV. Related Work

As early as the Apollo mission, information retrieval systems were developed to manage mission operations information. Those systems were limited by the techniques and computing power available at the time: a documentation system for guidance and navigation software and associated crew procedures had a simple indexing scheme to enable information retrieval[4]. More recent mission operations support systems have maintained reliance on simple indices and browsing. The Virtual Mission Operations Center (VMOC) primarily used commercial off-the-shelf information systems, such as Lotus Notes, to provide its information retrieval capabilities.[5] Guardian,[6] an information system for military operations, provides access to its collection through hierarchical index terms. The Collaborative Information Portal (CIP) facilitated navigation through NASA's Mars Rover Mission data products using metadata-based filtering and hierarchical file browsing.[7]

Our system focuses on retrieval of information using search keywords and metadata, as well as through cross-referencing and text similarity. Numerous automated information retrieval systems have supported searching of large document collections via keywords and metadata for decades (e.g., Medline[8] and LexisNexis). More recently, publishers of scientific literature have made the complete text of books and periodicals available for retrieval via keyword searches. The Science Citation Index (SCI), a commercial product, was developed after initial investigations proved the value of organizing and retrieving information based on cited prior work[9]. For many scientific collections, SCI provides the capability to perform the type of citation-guided search that XSearch provides for operations records. Originally, created manually, programmatic means were eventually developed to generate the SCI and similar citation indexes (e.g., CiteSEER, Google Scholar)[10, 11]. These methods have included probabilistic, statistical and lexical approaches to recognizing citation variations [12].

V. Discussion

A. Status and Future Directions

As of this writing, XSearch is in its final stages of testing prior to deployment in the MCC. Because cross-referencing and similarity detection are new concepts in the flight control environment, we plan to do a formal evaluation of these features to assess their utility. This assessment will likely be both qualitative and quantitative in nature. The qualitative evaluation will include interviews with flight control personnel to gather their opinions about XSearch and solicit additional requirements for search. We will also conduct comparative trials in which a set of information retrieval tasks is presented to users in two groups: those using XSearch, and those restricted to current tools. Performance measures will be compared across these groups. Quantitative analysis will focus on an examination of log files that XSearch generates as users interact with the system. These logs will be compared with search log files being generated by the chits, flight notes, and anomaly report systems currently used in MCC. We hope to detect any changes occurring in the types of searches being performed using XSearch vs. legacy tools, and will track user migration to our new system. The log files will also be used to assess user adoption of new features, such as cross-referencing and similarity.

We have already begun to develop several types of extensions to the baseline version of XSearch. First, we are moving beyond chits, flight notes, and anomalies to incorporate additional data sources to search. In particular, we are now working with several software change-tracking databases and a console log database used heavily by flight controllers. Second, we are developing the ability to detect new types of cross-reference patterns within the text of records, apart from chits, flight notes, and anomalies. Specifically, we are developing patterns to detect references to part numbers, problem reports, and software change records. Patterns for flight rules and flight procedures would also be highly desirable, but these pose additional challenges. The text references to flight rules and procedures are less standardized and more often ambiguous than those to chits, flight notes, and anomaly reports. Detecting these references may require application of more sophisticated parsing techniques. Even more challenging is the extraction of inferred references where explicit identifiers are not present, as discussed in Section II. Finally, we are generalizing our integration architecture to facilitate more rapid incorporation of new data sources and cross-referencing patterns.

Although we built XSearch to support space mission operations, there is nothing specific to this domain that would prevent its application elsewhere. We can imagine other applications of our context reconstruction approach in areas where text documents reference other documents or records. For example, in legal or governmental areas, documents are heavily cross-referenced using explicit identifiers referring to specific sections of the legal code, legal precedents, decisions, etc. Other possible application areas include engineering and building construction.

B. Implementation Challenges

Implementing XSearch required us to address several key challenges. Some of these challenges stem from performance constraints; others arise due to the complexity of the MCC workflow; yet others are the result of legacy data processing decisions that impact our ability to search across databases and compute cross-references:

- *Scaling and efficiency issues for similarity computation*: The three databases we analyzed for cross-references contain over 60,000 active records. The similarity computations are resource-intensive and must be performed for each pair of records, requiring more than 3.6 billion comparisons executing over 36 hours on an Intel® Xeon™ dual-core 3.05Ghz processor with 3.75 GB RAM. However, once these records have been processed, indexing of new records (which are produced at an estimated maximum rate of 40 records per hour) can be accomplished in an acceptable timeframe (i.e., under 10 minutes). Nevertheless, as the number of records in the overall corpus grows – both as a result of new record entry by flight controllers and by the addition of records from data sources newly-incorporated into XSearch – the processing time will increase as a function of the square of the number of records added. Ultimately, steps will be necessary to manage the impact of increased processing time on system performance.
- *Record modifications*: Chits, flight notes, and anomaly records go through a sequence of modifications by flight controllers as they negotiate content and wording changes over a period of several days to as much as a month or more. Each time the text of a record is modified, that record must be re-analyzed for cross-references and similarity. In some cases, references may have been deleted or added, so the XSearch cross-reference table must be modified correspondingly. When the text of a record changes, XSearch recomputes the set of similarity scores comparing the modified record with all others. But in fact, changes to similarity scores are not restricted to the scores in this set; any modification that introduces or eliminates words in the record's text can affect other similarity scores, as well, due to changes in the value of the IDF term in Eq. (1). Given the large size of the corpus, these impacts are typically minor, but in some cases significant scoring changes can occur. XSearch does not recalculate all similarity scores that may be affected by record modifications due to resource limitations; XSearch operating requirements mandate that updates are processed in near real-time. To mitigate the impact of inaccurate similarity scores, we plan to recompute all scores periodically using a background batch job.
- *Lack of unique flight note identifiers*: Although most of the record identifiers used by flight controllers to identify flight notes are unique, a fraction is not. Early on, flight teams recycled record identifiers periodically, although this practice was discontinued after a couple of years. When XSearch encounters a reference to a non-unique identifier, it displays all matching records. Resolving which record the user intended to reference is beyond the scope of our efforts.
- *Cross-database inconsistencies*: The structure of the chit, flight note, and anomaly report database schemas are similar, but not identical. XSearch formulates queries across these three database schemas and presents information in a unified manner. Sometimes, this requires mapping the terminology used in one database to the corresponding terminology used in another. For example, the table and column names that store a record's title may be different across the three databases. Additionally, it may be necessary to map the values stored in the table columns. For example, a record status of 'closed' in one database may correspond to the value 'finished' in another. As XSearch incorporates additional and more heterogeneous data sources, the problem of maintaining mappings will become even more important. We are developing new techniques that will enable XSearch to integrate new data sources with minimal changes to the XSearch source code.

C. Observations

NASA's information management systems for Shuttle and ISS mission operations are built primarily in a *record-centric* or *document-centric* fashion. Each type of record or document is isolated in its own database or repository and accessed using a separate system that handles data entry, modification, and search. Instead, we believe that these systems should be built in an *entity-centric* fashion, where operations entities – including events, activities, issues, people, and systems – form the central organizing components around which information is structured and accessed. These entities should in turn be connected to relevant documents and records. In current systems, entities are not represented explicitly and little metadata are captured to describe them. Yet users of these systems intuitively think in terms of entities, rather than records or documents. This causes an impedance mismatch between the user and the system, thereby inhibiting effective information access.

The lack of a common, underlying, entity-centered information model is also one of the main barriers to conducting integrated search across the patchwork of flight control resources in MCC. Without this type of connecting infrastructure, it is very difficult to relate data stored across resources. XSearch attempts to perform a kind of reconstructive surgery, piecing together bits and pieces of connecting information that is missing from existing resources; by its very nature, this process is incomplete and error-prone. Next generation aerospace information management systems must be built upon a solid common base to avoid this kind of disconnect across resources. NASA's Constellation Program is currently designing and implementing an information infrastructure consistent with this recommendation.

VI. Conclusion

In this paper, we have argued that locating and interpreting information is critical to flight controllers' work and ultimately to the safety of the astronauts aboard the ISS and Shuttle. Flight controllers must be able to respond to requests for information, troubleshoot anomalous situations, and plan for upcoming activities – all in a rapid and effective manner. XSearch attempts to eliminate some of the access and search barriers that impact the flight controller's ability to work effectively. In addition, by providing cross-references, the system aims to improve their understanding of search results by providing contextual information that enhances situational awareness and enables more informed decision-making.

Acknowledgments

We would like to thank our colleagues who have worked closely with us on the management, design, implementation, and deployment of XSearch: Chris Knight, Mohana Gurram, Ian Sturken, Steve Lin, Keith Swanson, David Hall, Lisa Faithorn, Cori Schauer, and Dave Korsmeyer from NASA Ames Research Center; Tim Hall, Brian Ulman, Troy LeBlanc, Chuck Schultz, James Hart, Jenny Yang, Jon Shipman, Jason Hadlock, Greg Whitney, Jason Toschlog, and Ernie Smith from NASA Johnson Space Center. This work was supported with funding from the NASA Constellation Program Mission Operations Project Office.

References

[1]Cowie, J. and Lehnert, W., "Information extraction," *Communications of the ACM*, vol. 1, Jan. 1996, pp. p. 80-91.

[2]Riloff, E. and Hollaar, L., "Text databases and information retrieval," *ACM Computing Surveys*, vol. 1, March 1996, pp. 133-5.

[3]Salton, G., "A blueprint for automatic indexing," *SIGIR Forum*, vol. 16, pp. Inclusive Pagination: p. 22-38.

[4]Dunbar, J. C., Larson, R. A. and Augart, P. T., "An Automated Documentation Technique for Integrating Apollo Crew Procedures and Computer Logic." http://klabs.org/history/history_docs/mit_docs/1719.pdf: Massachusetts Institute of Technology, 1966.

[5]Bane, R. and Fox, J. A., "The Design and Implementation of the VMOC Prototype," *The Fourth International Symposium on Space Mission Operations and Ground Data Systems (SpaceOps '96)*, 1996.

[6]Kenyon, H. S., "Guardian Watches Over Force Alignment," in *Signal*, October 2005.

[7]Mak, R. and Walton, J., "The Collaborative Information Portal and NASA's Mars Rover Mission," *IEEE Internet Computing*, vol. 9, pp. 20-26.

[8]Bachrach, C. A. and Charen, T., "Selection of MEDLINE contents, the development of its thesaurus, and the indexing process.," *Med Inf*, vol. 3, pp. 237-54.

[9]Price, D. J. d. S., "Networks of Scientific Papers," *Science*, vol. 149, pp. 510-5.

[10]Roth, D. L., "The emergence of competitors to the Science Citation Index and the Web of Science," *Current science.*, vol. 89, 2005, pp. 1531-1536.

[11]Jacso, P., "As we may search - Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases," *Current science.*, vol. 89, 2005, pp. 1537-1547.

[12]Lawrence, S., Giles, C. L. and Bollacker, K., "Digital libraries and autonomous citation indexing.," *IEEE Computer*, vol. 32, pp. 67-71.